Is indigenization in probabilistic constraints a sign of different grammars?

Insights from syntactic variation in New Englishes

Melanie Röthlisberger
melanie.rothlisberger@kuleuven.be

KULeuven
Quantitative Lexicology and Variational Linguistics

Introduction

Indigenization in World Englishes

"indigenization of language structure mostly occurs at the interface between grammar and lexis, affecting the syntactic behavior of certain lexical elements. Individual words, typically high frequency items, adopt characteristic but marked usage and complementation patterns."

(Schneider 2007: 46; emphasis mine)

(1) *no worries*

(2) *this hair-style is called* **as** *'duck tail'*

Probabilistic indigenization

"the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties. To the extent that patterns of variation in a new variety A [...] can be shown to differ from those of the mother variety, we can say that the new pattern represents a novel, if gradient, development in the grammar of A."

(Szmrecsanyi et al. 2016: 133)

Work plan

1. Explore the extent of indigenization in syntactic variation patterns across varieties of English by
   - . . . using the (comparative) variationist method
     (Labov 1982; Tagliamonte 2001)
   - . . . to study probabilistic factors constraining the alternation(s)
2. Explore the boundaries of indigenized grammars

The English dative alternation

(3) **ditransitive dative**

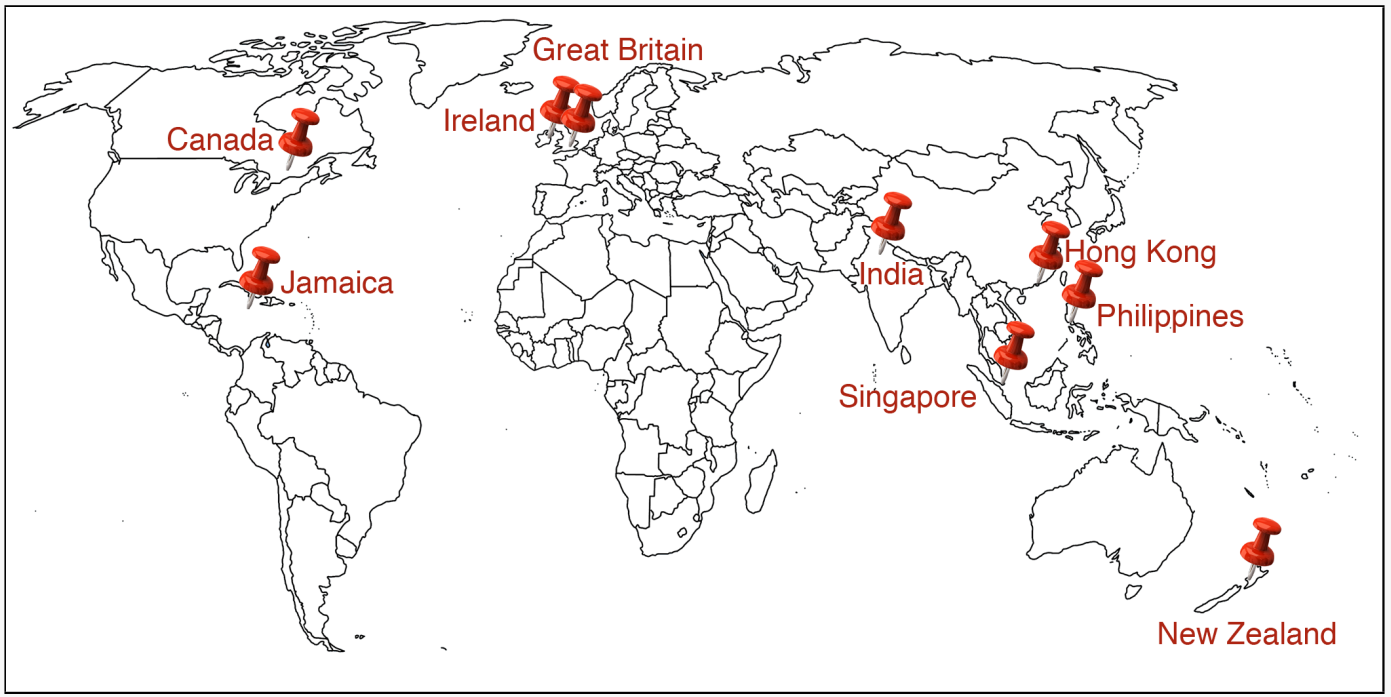
He gives [Mary]$_{recipient}$ [a present]$_{theme}$

(4) **prepositional dative**

He gives [a present]$_{theme}$ to [Mary]$_{recipient}$

Varieties of English covered

- British E, Canadian E, Indian E, Singapore E, Irish E, New Zealand E, Hong Kong E, Jamaican E, Philippines E



Data sources

- International Corpus of English (ICE) - series
- 60% spoken (transcriptions), 40% written texts
- 1m words per subcorpus
- 500 texts, 2,000 words per text
- 12 different registers, same corpus structure

Previous research

- statistical tendencies and processing principles underlying the dative alternation are shared across varieties
- stability in probabilistic grammars
  - 'easy' comes first $\rightarrow$ congruent effect
  - easy = animate, definite, pronominal, short
- variability (indigenization) in probabilistic grammars
  - recipient animacy: NZE vs. AmE
  - end-weight: AmE vs. AusE

  (e.g. Bresnan and Hay 2008; Bresnan and Ford 2010)
- shortcomings

Methodology

Dative tokens

(e.g. Bresnan et al. 2007)

1. retrieval of dative variants using verb list and perl script
2. restrict to choice context (incl. pronouns)
3. code for numerous (language-internal) factors: length (weight ratio), complexity, pronominality, givenness, definiteness, person, animacy, concreteness of theme, verb sense
4. code for language-external factors: Variety, register

N=8,549

Empirical investigation

- mixed-effects logistic regression
- full model: deviation coding for VARIETY and REGISTER: compare every level to the mean of ALL levels
- predicted outcome: prepositional dative
- `glmer()` function in Rs `lme4` package

  (Bates et al. 2015; Harrell 2015)
- random effects include
  - verb lemma and verb sense
  - corpus structure
  - recipient and theme head lemmas

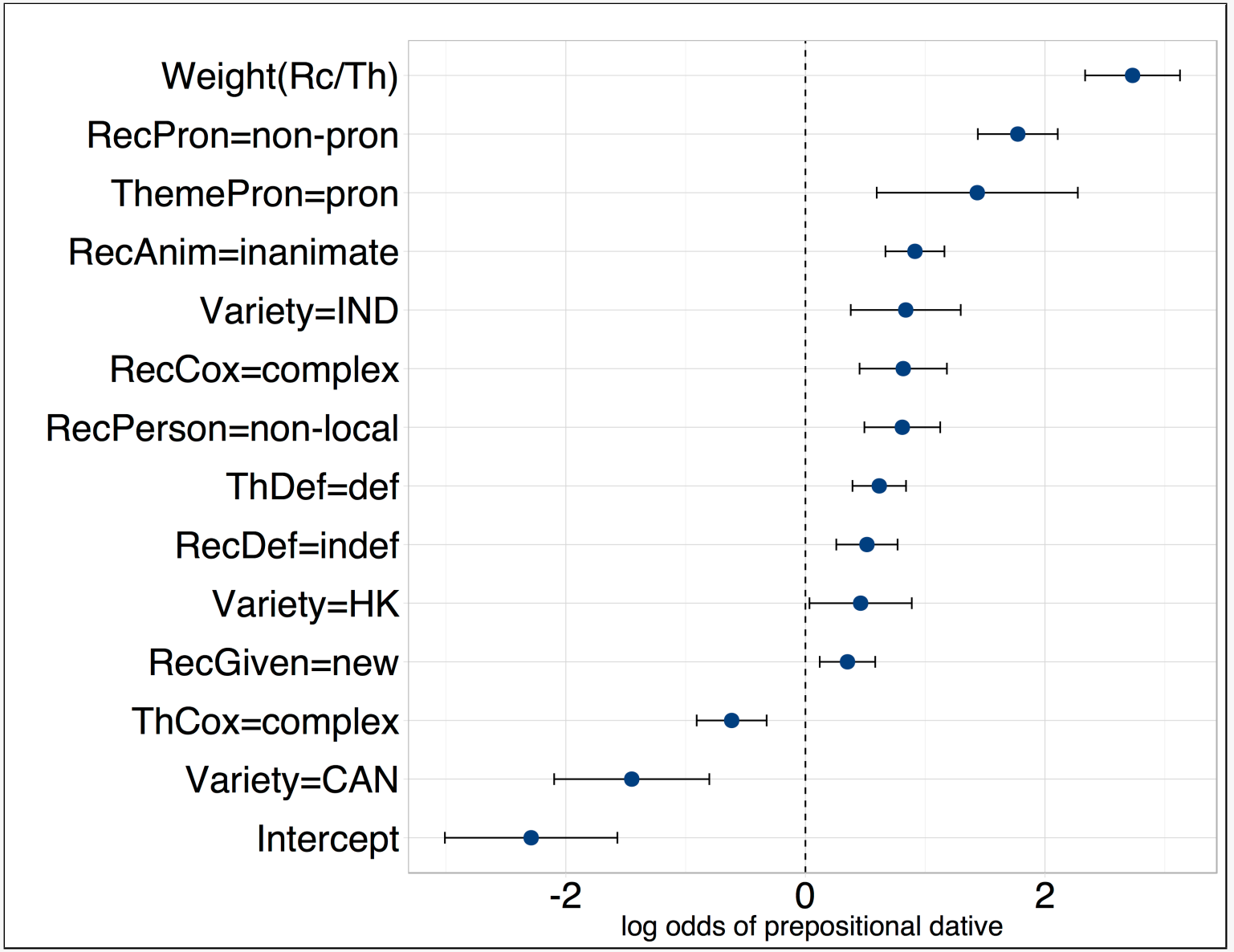
Results

The full dative model

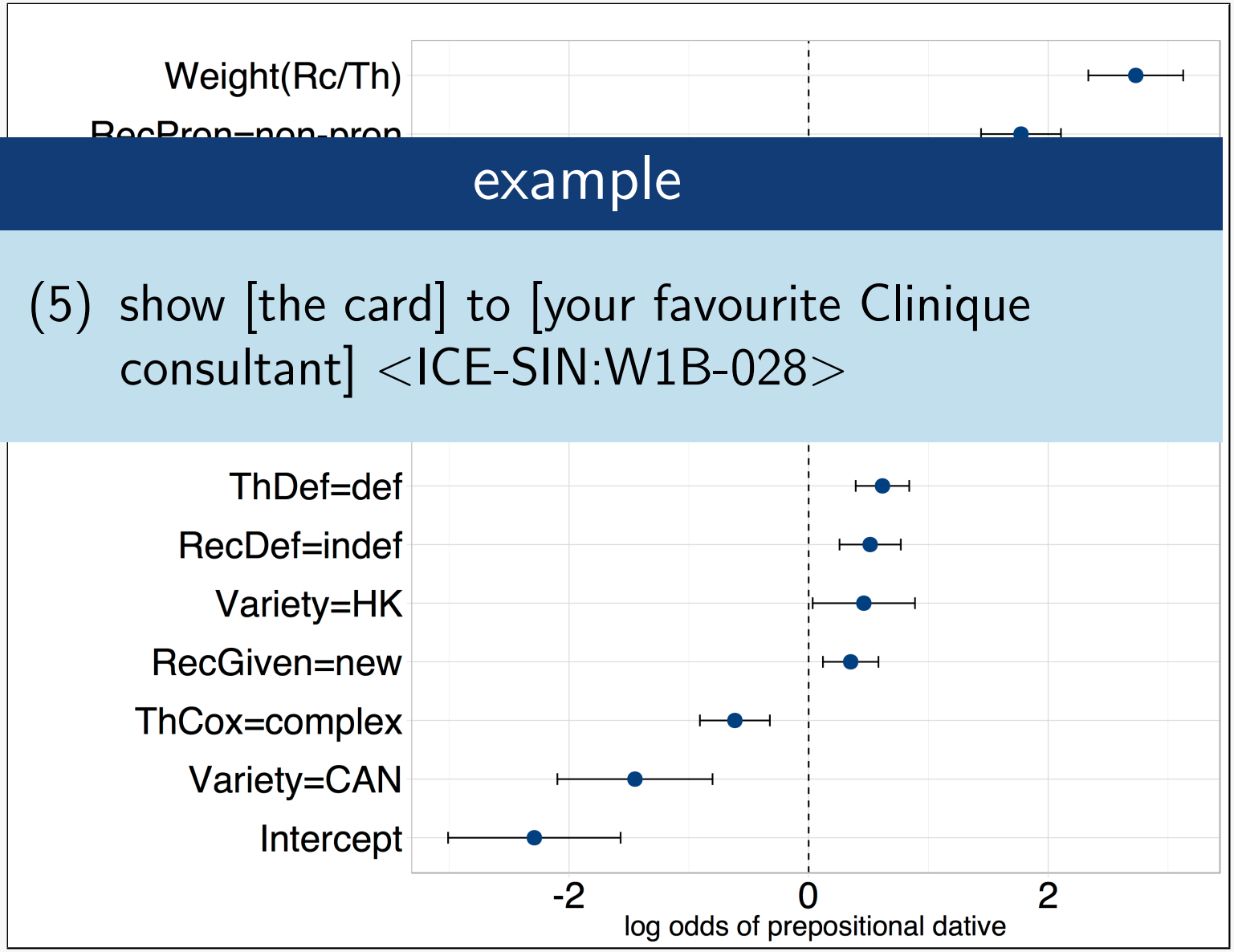
Response = {ditransitive, prepositional}

Response ~ (1|VerbLemma/VerbSense)
+ (1|ThemeHead)
+ (1|CorpusStructure)
+ RecComplexity
+ RecGivenness
+ ThemeComplexity
+ RecPerson
+ RecDefiniteness
+ ThemePron
+ RecAnimacy
+ ThemeGivenness
+ ThemeDefiniteness
+ Variety *
(Register + RecPron + ThemeConcreteness + WeightRatio)

Main effects



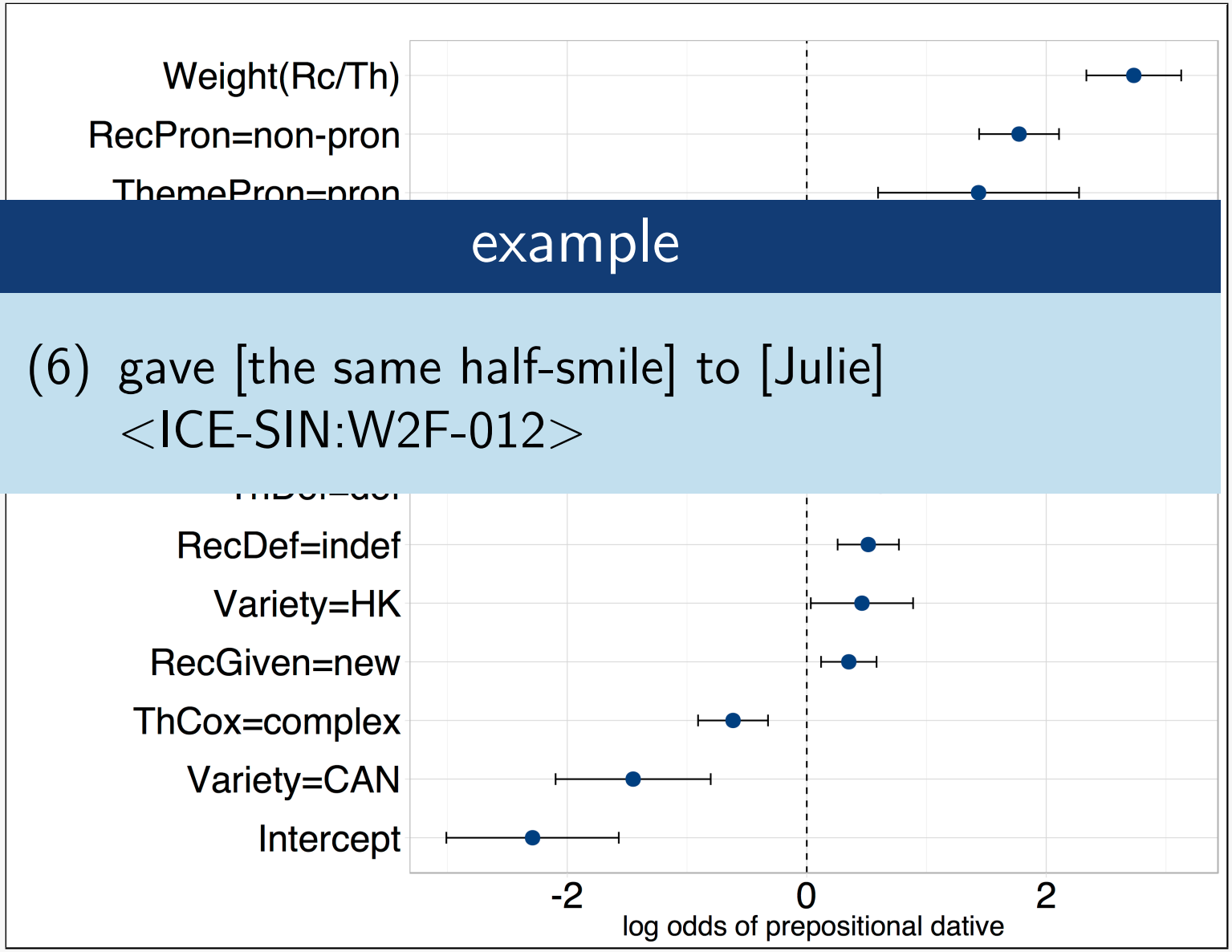
Main effects



example

(5) show [the card] to [your favourite Clinique consultant] <ICE-SIN:W1B-028>

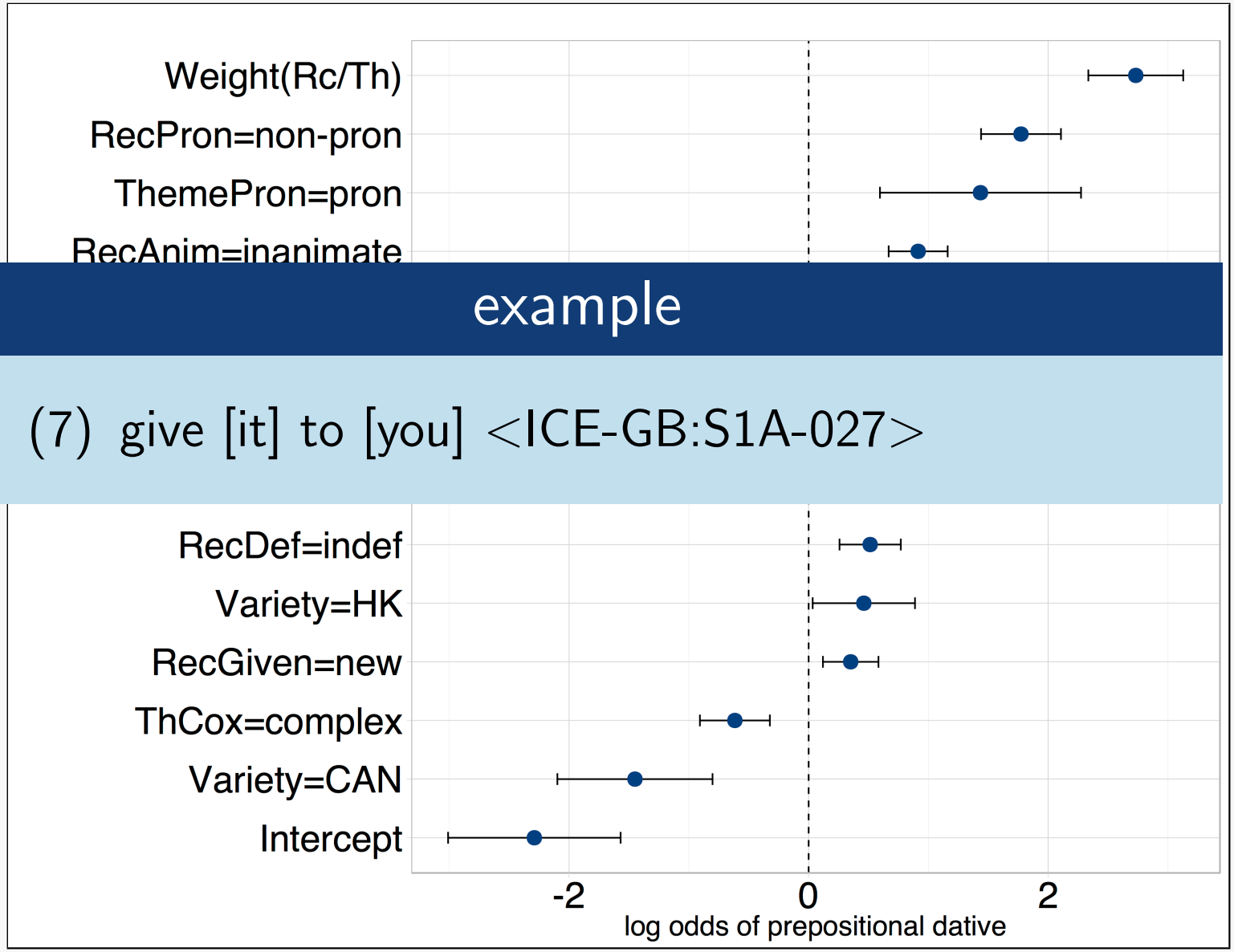
Main effects



example

(6) gave [the same half-smile] to [Julie]
<ICE-SIN:W2F-012>

Main effects



example

(7) give [it] to [you] <ICE-GB:S1A-027>

Main effects

- all predictors influence the choice of construction as predicted:
  - given >new
  - animate >inanimate
  - definite >indefinite
  - pron >non-pron
  - short >long

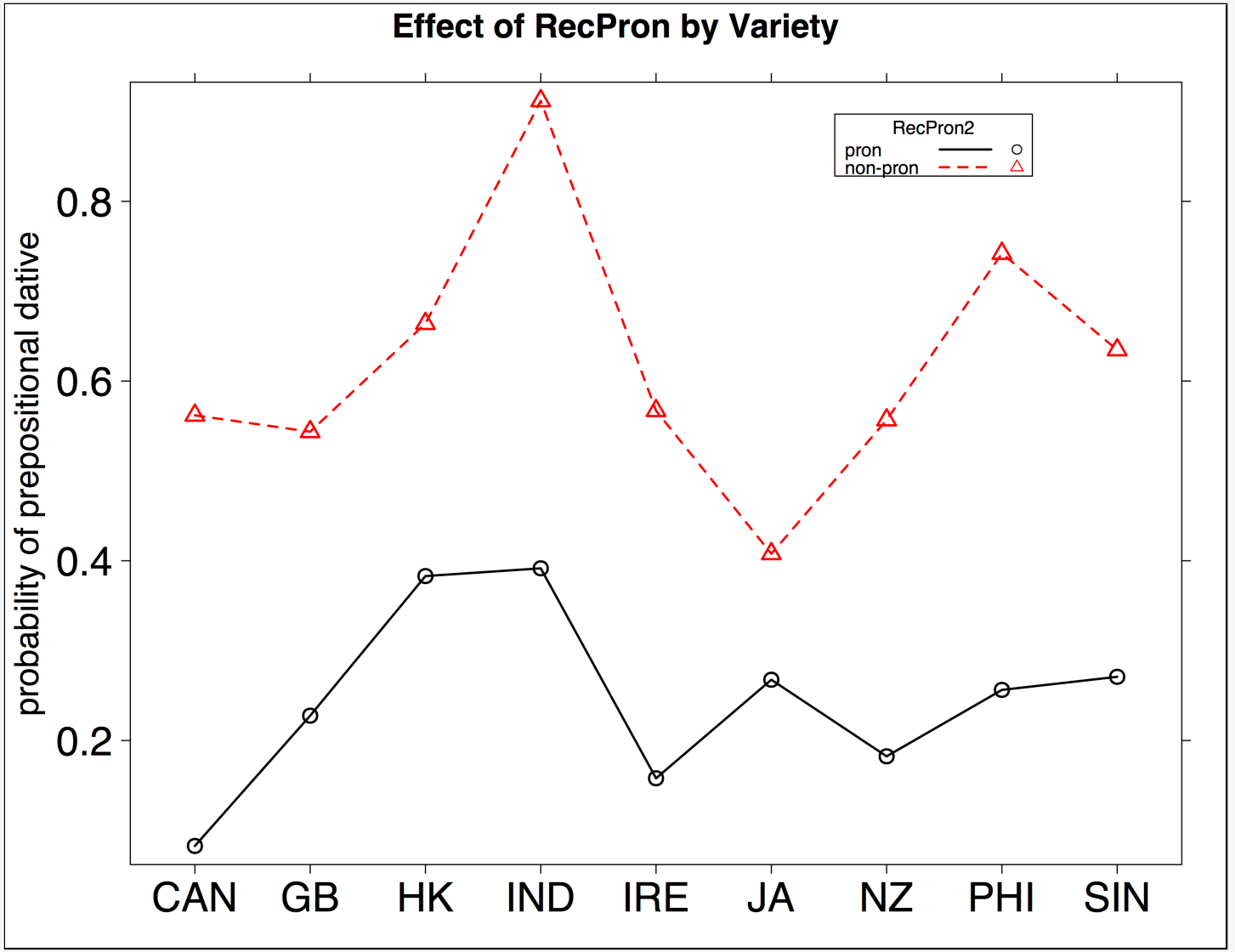
  recipient >theme $\rightarrow$ **ditransitive**
  theme >recipient $\rightarrow$ **prepositional**

The full dative model

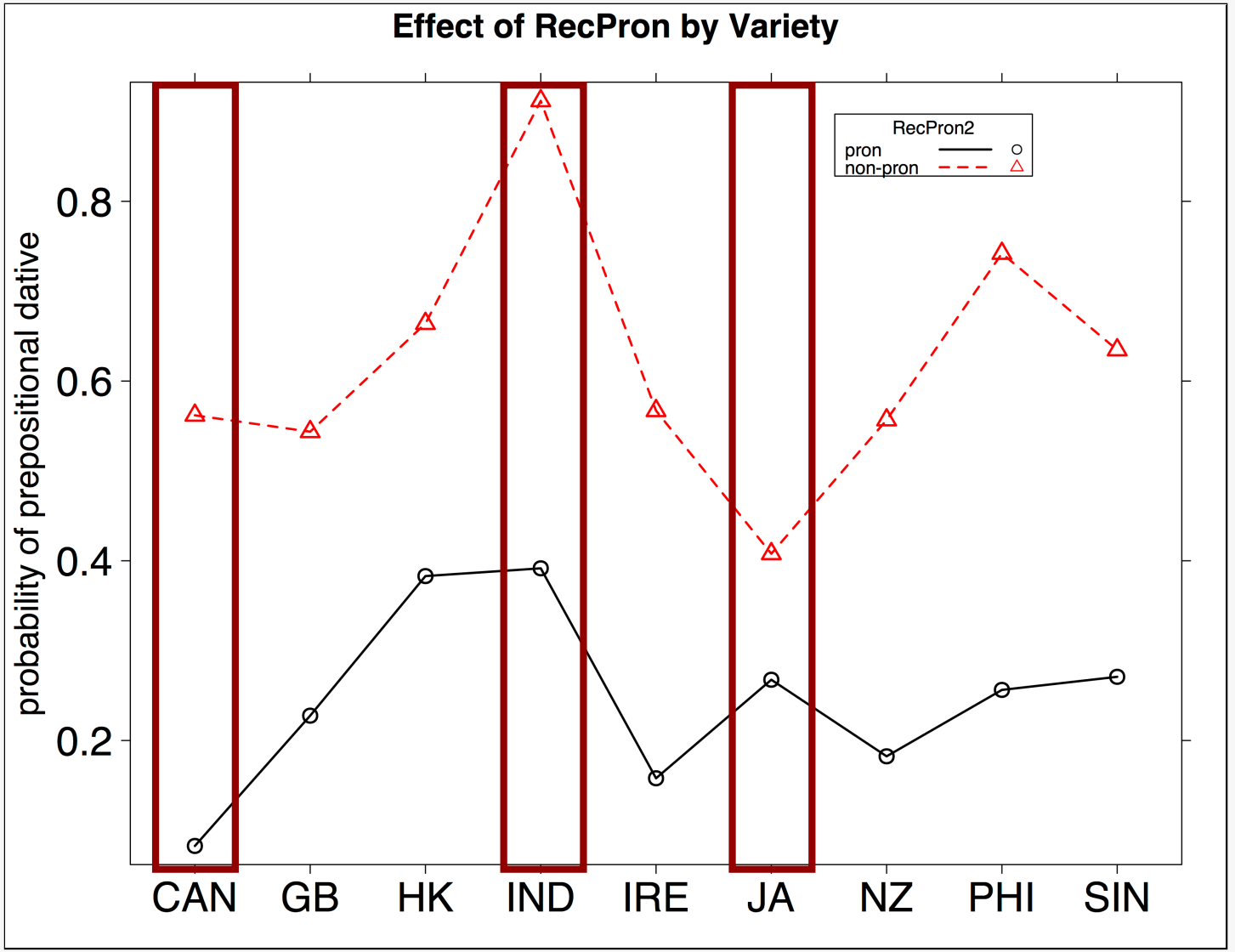
$$\text{Response} = \{\text{ditransitive, prepositional}\}$$

$$\begin{aligned}\text{Response} \sim\ & (1|\text{VerbLemma}/\text{VerbSense}) \\ & + (1|\text{ThemeHead}) \\ & + (1|\text{CorpusStructure}) \\ & + \text{RecComplexity} \\ & + \text{RecGivenness} \\ & + \text{ThemeComplexity} \\ & + \text{RecPerson} \\ & + \text{RecDefiniteness} \\ & + \text{ThemePron} \\ & + \text{RecAnimacy} \\ & + \text{ThemeGivenness} \\ & + \text{ThemeDefiniteness} \\ & + \text{Variety} * \\ & (\text{Register} + \text{RecPron} + \text{ThemeConcreteness} + \text{WeightRatio})\end{aligned}$$

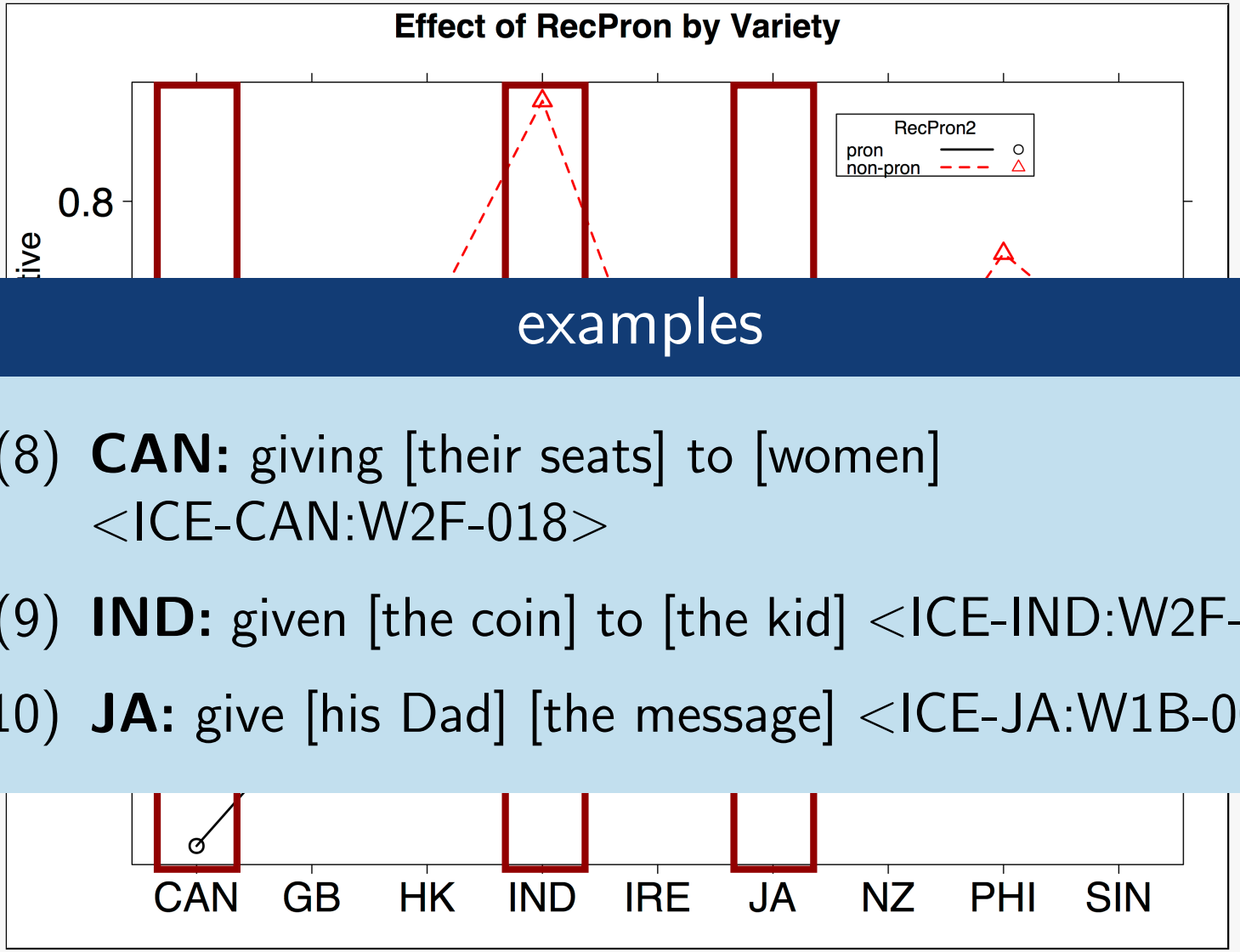
Interactions: RecPron



Interactions: RecPron



Interactions: RecPron



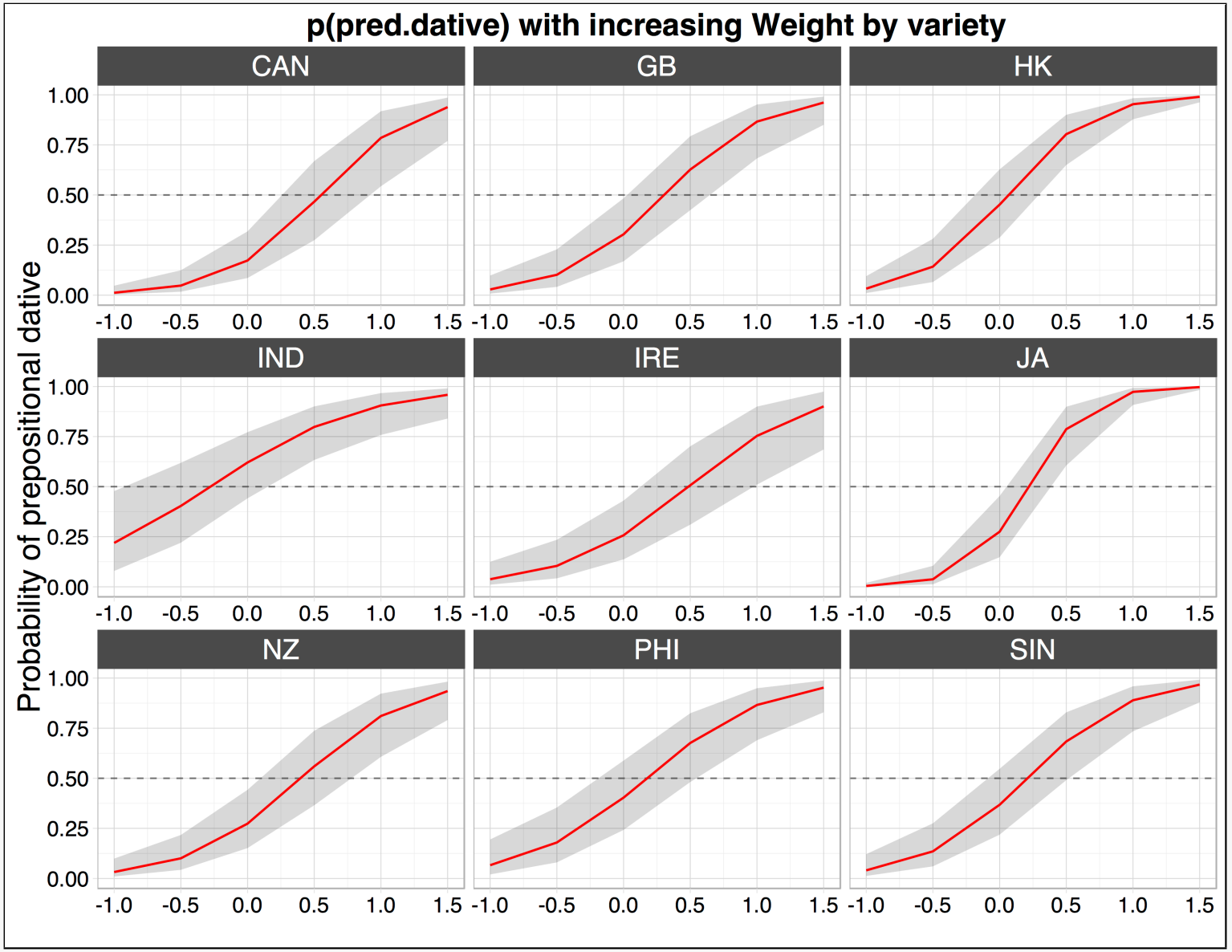
examples

(8) **CAN:** giving [their seats] to [women] <ICE-CAN:W2F-018>

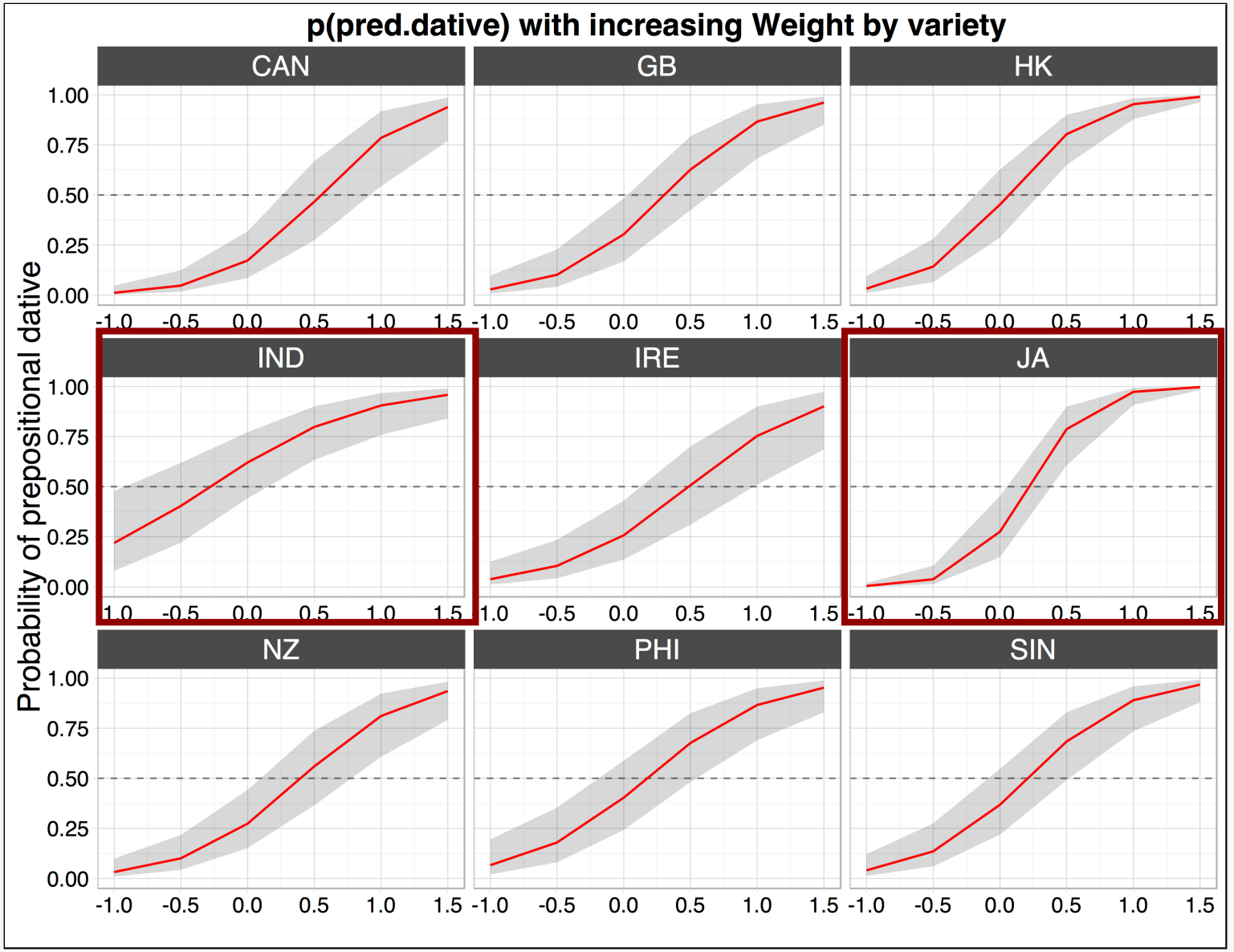
(9) **IND:** given [the coin] to [the kid] <ICE-IND:W2F-006>

(10) **JA:** give [his Dad] [the message] <ICE-JA:W1B-004>

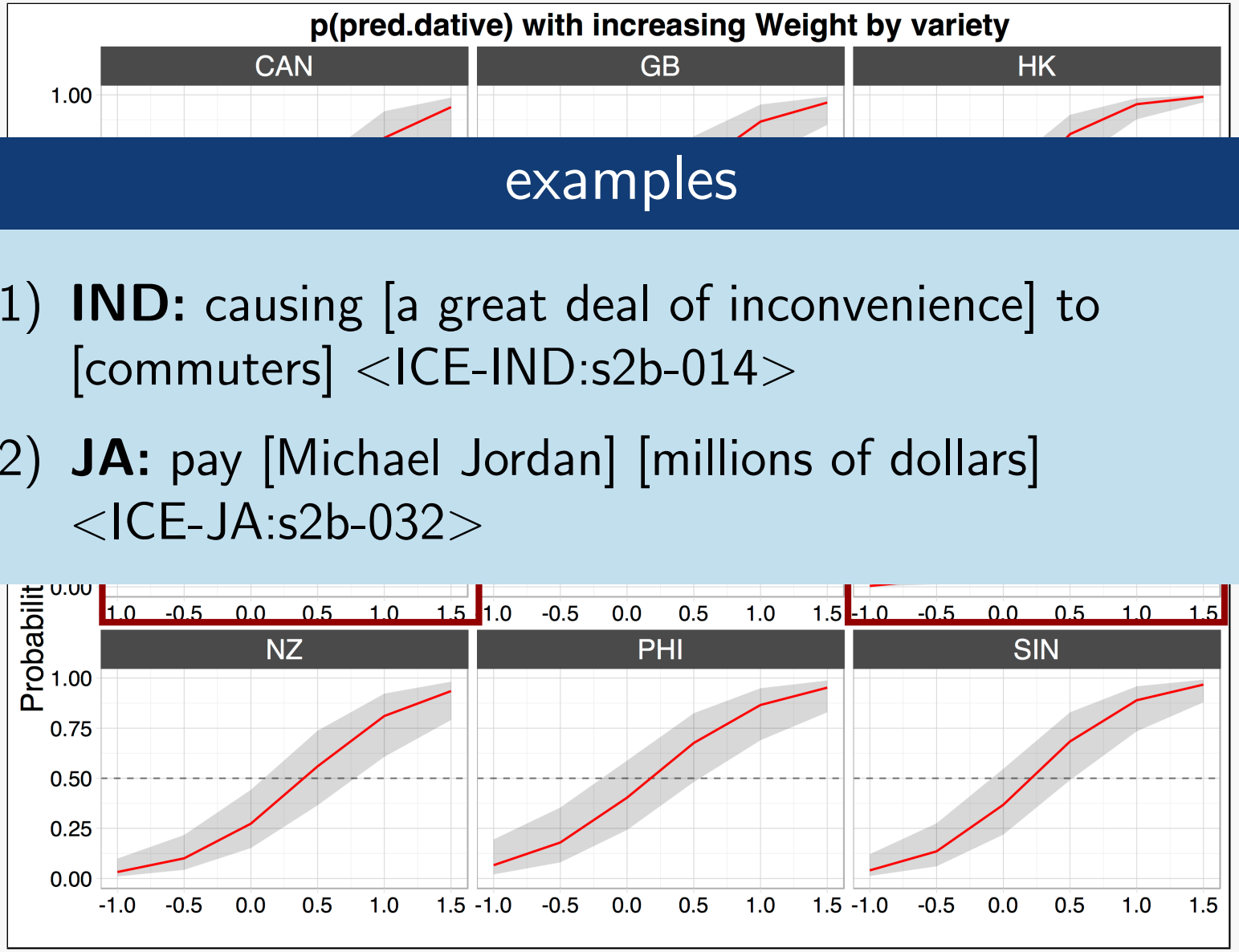
Interactions: Weight



Interactions: Weight



Interactions: Weight



examples

(11) **IND:** causing [a great deal of inconvenience] to [commuters] <ICE-IND:s2b-014>

(12) **JA:** pay [Michael Jordan] [millions of dollars] <ICE-JA:s2b-032>

Probabilistic indigenization

- recipient pronominality has a greater effect in Indian and Canadian English and a weaker effect in Jamaican English
- the effect of end-weight (short before long) is weaker in IndE and stronger in JamE (compared to all other varieties)

Two patterns

Syntactic variation in postcolonial Englishes is characterized both by qualitative stability regarding the choice of dative variant and localized (probabilistic) indigenization in the quantitative strength of individual constraints.

Discussion

Different grammars?

Is indigenization in constraints shaping syntactic variation a sign of different grammars?

- What is "different"?
- Where are the boundaries?
- Where do speakers deviate?

Finding boundaries between grammars

- **Explore deviations**: Under which circumstances does a language user make a different (dative) choice?
  ⇒ Multifactorial Prediction and Deviation Analysis with Regressions (MuPDAR)

  (see Gries and Adelman 2014; Gries and Deshors 2014, 2015)
- **Explore distances**: How (dis-)similar are grammars and where do we draw the line?
  ⇒ Multidimensional scaling

  (see Szmrecsanyi and Röthlisberger 2016; Szmrecsanyi 2010; Grafmiller forthcoming)

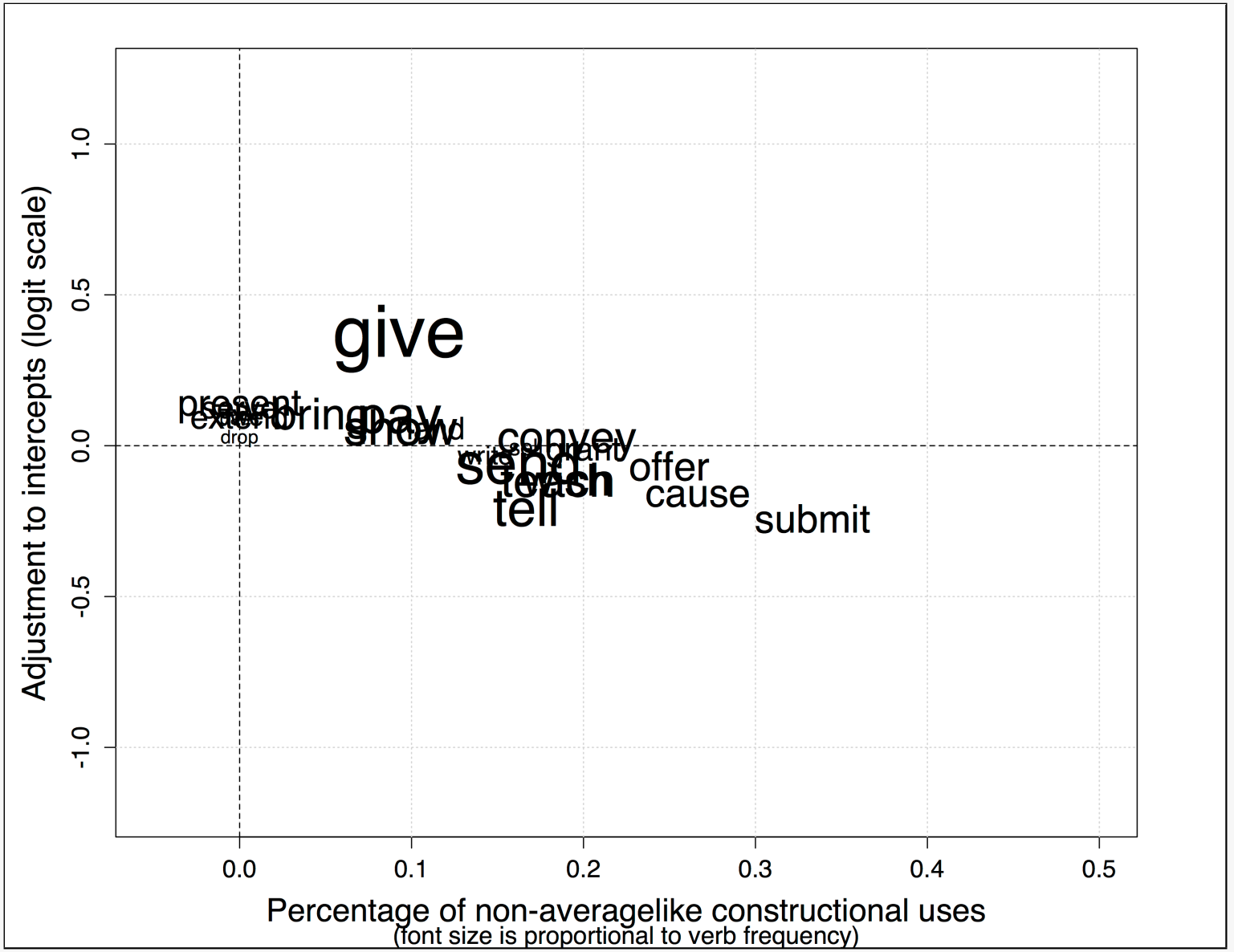
Exploring deviations: MuPDAR

1. create model R1 from full dataset without interactions
2. apply predictions from R1 (global average) to variety subset
3. compute "deviation"
   - 0: variety-specific choice = global choice
   - -0.5 to 0: variety-specific choice was ditransitive instead of pred.dative
   - 0 to 0.5: variety-specific choice was prepositional instead of ditransitive
4. create new factor 'SAME_CHOICE' as DV
5. run R2 on subset of the data (by-variety)

Context of deviation

- **Hot spots of "indigenization"**: Indian English, Jamaican English, Canadian English
- Conflict sites: in which context do speakers make a different choice?
- e.g. certain Verbs in IndE

Indian English: Verb complementation patterns



Finding boundaries between grammars

- **Explore deviations**: Under which circumstances does a language user make a different (dative) choice?
  ⇒ Multifactorial Prediction and Deviation Analysis with Regressions (MuPDAR)

  (see Gries and Adelman 2014; Gries and Deshors 2014, 2015)
- **Explore distances**: How (dis-)similar are grammars and where do we draw the line?
  ⇒ Multidimensional scaling of distance measures

  (see Szmrecsanyi and Röthlisberger 2016; Szmrecsanyi and Hinrichs 2008; Szmrecsanyi 2010; Grafmiller ming)

Measuring distances: MDS

Cooking recipe:

1. run 1 regression per variety (grammar)
2. use coefficient estimates to compute Euclidean distance matrix

	GB	CAN	NZ	IRE	JA	SIN	HK	PHI
CAN	4.410758							
NZ	4.137141	4.809944						
IRE	4.919131	4.358283	4.149647					
JA	5.467474	5.461546	4.044327	6.205969				
SIN	1.921338	4.082965	3.366878	4.685010	5.106863			
HK	2.856351	4.889179	3.012393	5.269595	4.847752	3.195134		
PHI	4.663174	4.551190	2.995564	3.857443	5.350859	3.695588	4.491004	
IND	3.700839	5.206690	2.759111	3.835607	5.459240	3.791418	2.922903	3.982615

Measuring distances: MDS

Cooking recipe:

1. run 1 regression per variety (grammar)
2. use coefficient estimates to compute Euclidean distance matrix
3. Explore patterns in variation:
   - **cluster analysis (Ward)**: which varieties behave similarly...?
   - **MDS**: ...in a three dimensional space?

Measuring distances: MDS

pipeline:

1 regression model / variety
$\Rightarrow$ 9 $\times$ 20 coefficient matrix
$\Rightarrow$ Euclidean distance matrix
$\Rightarrow$ cluster analysis (Ward)
$\Rightarrow$ MDS

Conclusion

- indigenization (also) occurs on the very subtle level of speakers' probabilistic grammar
- speakers' probabilistic grammar differs with regard to the strength of predictors
- dissimilarities between speakers' probabilistic grammar can be explored using statistical tools at hand
- indications of divergence in "probabilistic grammars"
- More features (syntactic, lexical) need to be added to find "isoglosses" between speakers' probabilistic grammar(s)

Thank you!

melanie.rothlisberger@kuleuven.be

http://wwwling.arts.kuleuven.be/qlvl/ProbGrammarEnglish.html

References I

Bates, D., M. Mächler, B. M. Bolker, and S. Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software 67*(1). LIB MR.

Bresnan, J., A. Cueni, T. Nikitina, and H. Baayen (2007). Predicting the Dative Alternation. pp. 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, J. and M. Ford (2010). Predicting Syntax: Processing dative constructions in American and Australian Varieties of English. *Language 86*(1), 168–213.

Bresnan, J. and J. Hay (2008, February). Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua 118*(2), 245–259.

Grafmiller, J. (forthcoming). Mapping out particle placement in varieties of English across the globe.

Gries, S. and A. Adelman (2014). Subject realization in Japanese conversation by native and non-native speakers: exemplifying a new paradigm for learner corpus research. In J. Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics 2014: New empirical and theoretical paradigms*, pp. 35–54. Cham: Springer.

Gries, S. T. and S. C. Deshors (2014, May). Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora 9*(1), 109–136.

Gries, S. T. and S. C. Deshors (2015). EFL and/vs. ESL? A multi-level regression modeling perspective on bridging the paradigm gap. *International Journal of Learner Corpus Research 1*(1), 130–159.

Harrell, F. E. J. (2015). Regression modeling strategies. LIB MR.

Labov, W. (1982). Building on empirical foundations. In W. Lehmann and Y. Malkiel (Eds.), *Perspectives on Historical Linguistics*, pp. 17–92. Amsterdam, Philadelphia: Benjamins.

Macdonald, M. C. (2013, January). How language production shapes language form and comprehension. *Frontiers in psychology 4*(April), 226. LIB MR.

Schneider, E. (2007). *Postcolonial English: Varieties Around the World*. Cambridge: Cambridge University Press. LIB MR.

References II

Szmrecsanyi, B. (2010). The English genitive alternation in a cognitive sociolinguistics perspective. In D. Geeraerts, G. Kristiansen, and Y. Peirsman (Eds.), *Advances in Cognitive Sociolinguistics*, pp. 141–166. Berlin/New York: Mouton de Gruyter. LIB MR.

Szmrecsanyi, B., J. Grafmiller, B. Heller, and M. Röthlisberger (2016). Around the world in three alternations: modeling syntactic variation in varieties of English. *English World-Wide 37*(2). LIB MR.

Szmrecsanyi, B. and L. Hinrichs (2008). Probabilistic determinants of genitive variation in spoken and written English: a multivariate analysis across time, space, and genres. In T. Nevalainen, I. Taavitsainen, P. Pahta, and M. Korhonen (Eds.), *The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present*, pp. 291–309. Amsterdam/Philadelphia: John Benjamins Publishing Company. LIB MR.

Szmrecsanyi, B. and M. Röthlisberger (2016, June). Context matters: The probabilistic grammar of international varieties of English.

Tagliamonte, S. (2001). Comparative sociolinguistics. In J. Chambers, P. Trudgill, and N. Schilling-Estes (Eds.), *Handbook of Language Variation and Change*, pp. 729–763. Malden and Oxford: Blackwell. LIB MR.